A Survey of the State of Explainable AI for Natural Language Processing

Year: 2020 | Citations: 426 | Authors: Marina Danilevsky, Kun Qian, R. Aharonov, Yannis Katsis, B. Kawas

---

Abstract

Recent years have seen important advances in the quality of state-of-the-art models, but this has come at the expense of models becoming less interpretable. This survey presents an overview of the current state of Explainable AI (XAI), considered within the domain of Natural Language Processing (NLP). We discuss the main categorization of explanations, as well as the various ways explanations can be arrived at and visualized. We detail the operations and explainability techniques currently available for generating explanations for NLP model predictions, to serve as a resource for model developers in the community. Finally, we point out the current gaps and encourage directions for future work in this important research area.